GB00/489

The Patent Office

EJV

INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP9 1RH

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed [signature]

Dated 8/4/99

Patents Form 1/77

The Patent Office

16FEB99 E425726-1 D01463
P01/7700 0.00 - 9903451.4

Patents Act 1977
(Rule 16)

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

1. Your reference: 30990053 UK

2. Patent application number *(The Patent Office will fill in this part)*: **9903451.4** 16 FEB 1999

3. Full name, address and postcode of the or of each applicant *(underline all surnames)*:
Hewlett-Packard Company
3000 Hanover Street
Palo Alto
California 94304
United States of America

Patents ADP number *(if you know it)*: 4965880001

If the applicant is a corporate body, give the country/state of its incorporation: Delaware, USA

4. Title of the invention: Similarity searching for documents

5. Name of your agent *(if you have one)*: Richard Anthony Lawrence

"Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*:
Hewlett-Packard Limited
IP Section
Filton Road
Stoke Gifford
Bristol BS34 8QZ

Patents ADP number *(if you know it)*:

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number

Country	Priority application number *(if you know it)*	Date of filing *(day / month / year)*

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application	Date of filing *(day / month / year)*

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:*
a) any applicant named in part 3 is not an inventor, or
b) there is an inventor who is not named as an applicant, or
c) any named applicant is a corporate body.
See note (d)): YES

Patents Form 1/77

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form	
Description	6
Claim(s)	3
Abstract	1
Drawing(s)	4

10. If you are also filing any of the following, state how many against each item.

Priority documents	
Translations of priority documents	
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	
Request for preliminary examination and search *(Patents Form 9/77)*	
Request for substantive examination *(Patents Form 10/77)*	
Any other documents *(please specify)*	

11. I/We request the grant of a patent on the basis of this application.

Signature [signature] Date 16/2/99

Richard Anthony Lawrence

12. Name and daytime telephone number of person to contact in the United Kingdom Janet Smith 0117-922-8026

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.

b) Write your answers in capital letters using black ink or you may type them.

c) If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.

d) If you have answered 'Yes' Patents Form 7/77 will need to be filed.

e) Once you have filled in the form you must remember to sign and date it.

f) For details of the fee and ways to pay please contact the Patent Office.

SIMILARITY SEARCHING FOR DOCUMENTS

30990053 UK

Field of Invention

The present invention relates to a method and means for searching to find similar documents in response to a query. The invention is particularly relevant to the use of one document as a query for a search to obtain similar documents.

Description of Prior Art

Similarity searching in databases of electronically stored documents is an important area of practical application. Such searching is well known for text. Typically, the input for such searching would be a text string, and the engine would then search the database matching entries against the text string and return entries with an acceptable similarity threshold. Similar searching is available for images.

Existing techniques are effective when the query is of essentially one data type: a text string only, or an image only. In general, however, an electronic document will consist of a combination a number of data types: a typical document might contain one or more text passages, one or more images, and line art. The text passages may also be readily subdividable into different types, such as headings, legends, and bulk text. Using existing techniques as indicated above, similarity searching will involve extraction of one element in a particular data type followed by similarity searching appropriate to that data type.

It is desirable to provide methods of similarity searching which allow the features of the document to be used appropriately in a search which is properly representative of the full document.

Summary of Invention

Accordingly, in a first aspect the invention provides method of searching a database to find documents similar to a query document, comprising: decomposing the query

document into elements of different data types; for one or more of the elements in a first data type, conducting a first data type similarity search to return match results from the database for the one or more elements in the first data type; for one or more of the elements in a second data type, conducting a second data type similarity search to return match results from the database for the one or more elements in the first data type; combining the match results from the first data type similarity search and the second data type similarity search to provide query document match results.

In a second aspect, the invention provides method of searching a database to find documents similar to a query document, comprising: decomposing the query document into elements of different data types; determining a layout element in a layout datatype from the spatial arrangement of the elements in the document; for the layout element, conducting a layout similarity search to return match results from the database for the layout element.

Brief Description of Figures

Specific embodiments of the invention are described below, by way of example, with reference to the accompanying drawings, of which:

Figure 1 shows a typical document page containing different data types;

Figure 2 shows steps in a method according to an embodiment of a first aspect of the invention for conducting a similarity search for the document shown in Figure 1;

Figure 3 shows the representation of the document shown in Figure 1 as a layout of datatypes, and indicates a search step usable in a further embodiment of the method of the invention; and

Figure 4 shows steps in a method according to an embodiment of the second aspect of the invention for conducting a similarity search for layout information.

Description of Embodiments

A typical document contains a plurality of data types. The most basic data types are text and images. Document 1 shown in Figure 1 contains a text block 12 - this text block is data in a first data type. Document 1 also contains two different kinds of image. One kind, image block 13, is a photographic image, typically consisting of an array of pixels in which each pixel has a colour value. The other kind, line art block 11, is also an image but a "drawn" one, readily representable as a combination of geometric or formulaic elements - and as such, typically readily scalable. Photographic images and line art images (hereafter "pictures" and "graphics") respond differently to different image processing and analysis techniques, and are most effectively treated as different data types. Moreover, pictures and graphics will generally serve a different purpose in a document, so it is also practical for the purpose of similarity searching to treat pictures and graphics separately.

The steps involved in similarity searching for the document of Figure 1 according to an embodiment of the first aspect of the invention are shown in Figure 2.

Firstly the document 1 is selected 21. For an electronic document, this could be achieved through any appropriate application capable of supporting the file type or file types of the document. For a physical document, this could be achieved by scanning the document using a scanner.

Secondly, the document is decomposed into separate elements: in the case of document 1, these elements are graphic block 11, text block 12, and picture block 13. In the case of text block 12, it is desirable for optical character recognition to be carried out at this point so that the text block element resulting from decomposition consists of ASCII text. Decomposition of the document is achieved by an analysis and recognition process through which the different parts of the document are recognised as being text, pictures or graphics. Decomposition of a document into separate data types in this way is known, using for example techniques identified in "Block Segmentation and Text Extraction in Mixed Text/Image Documents" by FM Wahl, KY Wong and RG Casey, Computer Graphics and Image Processing, Vol 20 (1982). Software adapted for use with proprietary scanners to decompose the elements of a scanned page into separate data types (in order to optimise the scanning process for each data type) is provided by Hewlett-Packard Company as "HP

PrecisionScan". The output of HP PrecisionScan is a set of elements each in a single data type, each of which can be selected for further processing.

The result of decomposition is a set of elements, each element having a single data type. For a particular data type, such as text, then either all text is determined to be part of a single element, or else physically distinct areas of text are considered as separate elements, depending on how the decomposition is carried out. In one version of the embodiment all the elements of the document are used in similarity searching: in other versions one or more of the elements are selected for use in similarity searching (or the user is even allowed an opportunity to select part of an element for such further processing).

Separate elements are then used in similarity searching 23 against a database, for example a database representing content available on the World Wide Web. Should all the elements be of one data type, this reduces to a conventional similarity searching problem addressable with a single search engine for the relevant data type. However, if elements are of different data types, then separate search engines are used for each data type. Appropriate search engines for similarity searching for different data types are known. For example, for text, appropriate linguistic matching toolkits are available from Teragram and Inxight.

The result of the similarity searching is a set of series of matching scores for documents in the database, such a set existing for each element searched. Each of these search scores need to be normalised 25 for combination 26 to achieve a combined search result 27. The normalisation step 25 is to ensure that a correct balance is given to the results of the different searching steps 23. This can either be to weight each element of the document equally, to weight each element of the document according to its perceived importance in the document, or according to a user assessment of the relative importance of the different elements of the document. The combined result 27 is as for conventional similarity searching: a series of matching scores (generally expressed as percentages) listing documents in the database from best towards worst matches.

Further use can be made of information derived from page decomposition in similarity searching. In addition to the separate elements provided by page decomposition (graphic 11, text block 12, and picture 13), further information is provided in the arrangement of the different elements within the document. As is shown in Figure 3, a further output available from page decomposition is a data type plan 31 representing the document as a line art block, a text block, and an image block, arranged vertically in sequence: this data type plan can itself be used as a layout data type. This allows yet another element - the layout data type element - to be used in searching 32 of a database (provided that layout information is available in or derivable from the database entries). The results of similarity searching for such a layout element can be combined with similarity searches for other elements exactly as described in Figure 2.

In an embodiment according to the second aspect of the invention, similarity searching is conducted using the layout data type alone. The steps to be followed are essentially as in conventional similarity searching - this is shown in Figure 4, with elements common to the first aspect of the invention given the same reference numbers as in Figure 2. Layout similarity searching, whether used on its own or as one of the elements in a combined search as described in the first aspect of the invention, is more powerful if a number of different data types are used for text and for overall document type. Using a rule based approach, different text blocks and whole documents, especially in the case of formal workflow documents, can be assigned particular functions with relatively high confidence. For example, it is well known that isolated text blocks at the top of a page and handwriting at the bottom are suggestive of a letter, and so different spatial regions of the document can be assigned to appropriate functional fields (address, letter text etc) - likewise, table and currency totals in a document can be identified as a discrete element, and their presence limits the document to another group (bill, quote or invoice). Layout searching can thus involve matching to templates representing different workflow document types (thus promoting matching of a document determined to be a letter against other letters). An appropriate mechanism is to normalise a layout for size, orientation and skew, and then carrying out an "exclusive or" operation on the query element and the layout records in the database - this will be effective provided that all records involved have a broadly common format.

The skilled man will appreciate that modifications of the embodiments described above can readily be carried out without departing from the invention as defined in the claims.

CLAIMS

30990053 UK

1. A method of searching a database to find documents similar to a query document, comprising:

decomposing the query document into elements of different data types;

for one or more of the elements in a first data type, conducting a first data type similarity search to return match results from the database for the one or more elements in the first data type;

for one or more of the elements in a second data type, conducting a second data type similarity search to return match results from the database for the one or more elements in the first data type;

combining the match results from the first data type similarity search and the second data type similarity search to provide query document match results.

2. A method as claimed in claim 1, wherein one of the data types is representative of text.

3. A method as claimed in claim 2, wherein a plurality of the data types are representative of text, separate data types of the plurality being representative of different functional blocks of text.

4. A method as claimed in any preceding claim, wherein one of the data types is representative of pictorial images.

5. A method as claimed in any preceding claim, wherein one of the data types is representative of graphical images.

6. A method as claimed in any preceding claim, wherein one of the data types is representative of the arrangement of other data types within the document.

7. A method as claimed in any preceding claim, wherein the step of similarity searching to return match results is carried out, separately, for a plurality of elements having between them more than two data types.

8. A method as claimed in any preceding claim, where all features of a common data type in the document are treated as one element.

9. A method as claimed in any of claims 1 to 7, where spatially distinct features of a common data type in the document are treated as separate elements.

10. A method as claimed in any preceding claim, wherein elements are user selectable or deselectable for the step of similarity searching.

11. A method as claimed in any preceding claim, wherein the similarity searching results for separate elements are weighted before combination.

12. A method as claimed in claim 11, wherein said weighting is user selected.

13. A method as claimed in claim 11, wherein said weighting is attributed according to a determined significance of each relevant element in the document.

14. A method of searching a database to find documents similar to a query document, comprising:

decomposing the query document into elements of different data types;

determining a layout element in a layout datatype from the spatial arrangement of the elements in the document;

for the layout element, conducting a layout similarity search to return match results from the database for the layout element.

15. A method as claimed in claim 14, wherein the layout similarity search involves searching against templates representative of different document types.

16. A method as claimed in claim 14, wherein the elements include elements of separate data types representative of different functional blocks of text.

17. A method as claimed in claim 14 or claim 16, wherein the elements include elements of data types representative of images.

ABSTRACT

Similarity Searching For Documents

30990053 UK

A method of searching a database, such as a database representative of content on the World Wide Web, to find documents similar to a query document, involves a step of decomposing 22 the query document 1 into elements of different data types. After this, for one or more of the elements in a first data type, a first data type similarity search is conducted 23 to return match results from the database for the one or more elements in the first data type. For one or more of the elements in a second data type, a second data type similarity search is conducted to return match results from the database for the one or more elements in the first data type. The match results from the different data types are combined with an appropriate weighting to provide query document match results. Data types can include text, picture and graphics, and also the layout of the overall document.

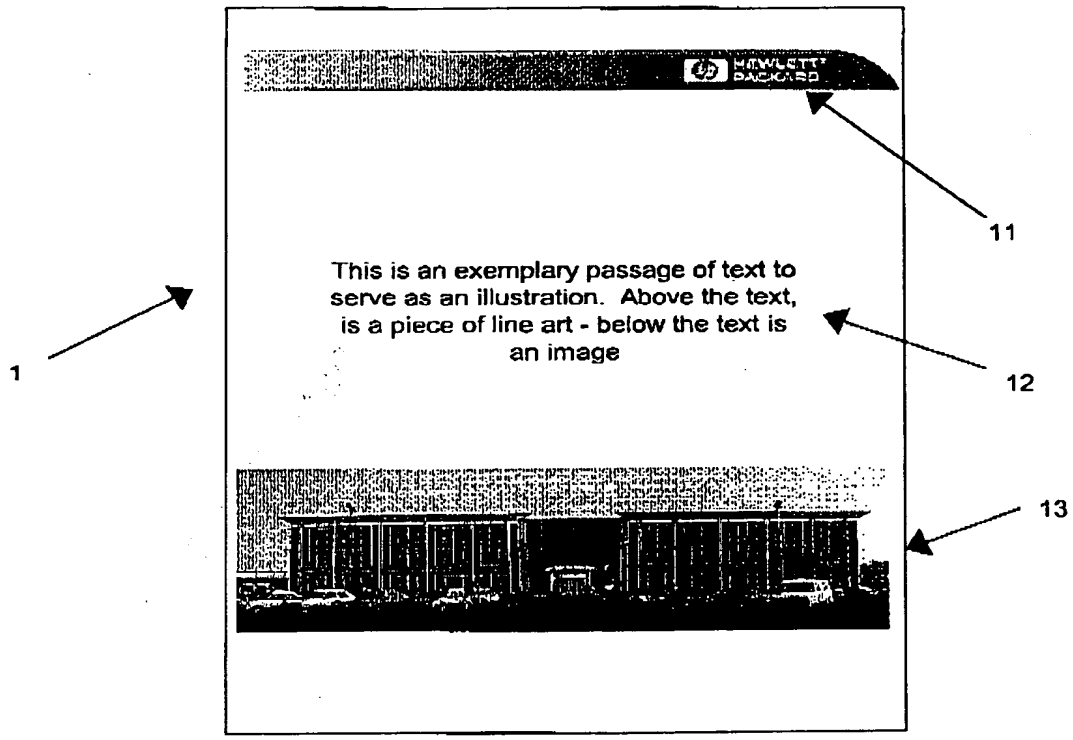
This is an exemplary passage of text to serve as an illustration. Above the text, is a piece of line art - below the text is an image
1
11
12
13

Figure 1

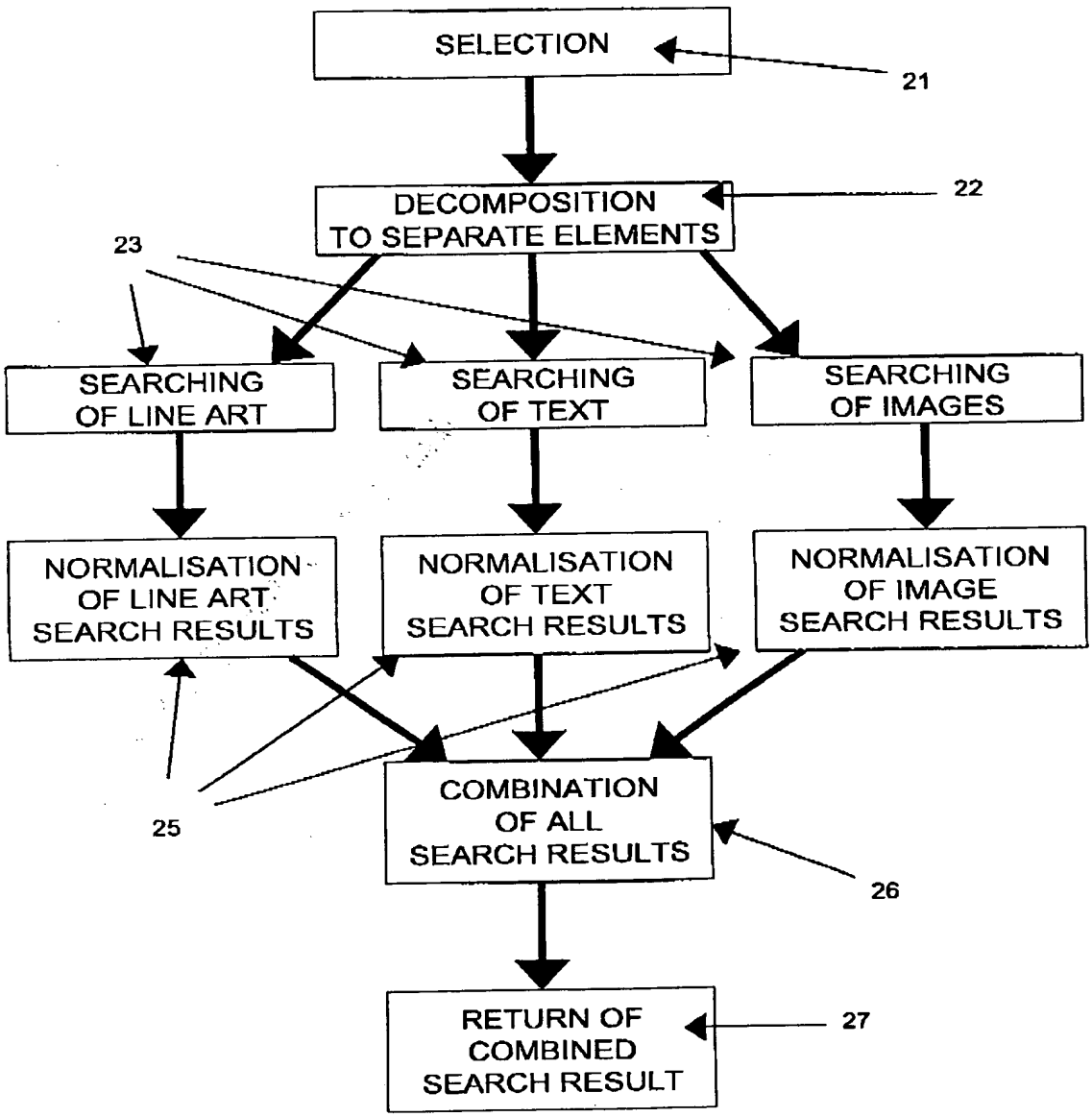
SELECTION
21
DECOMPOSITION TO SEPARATE ELEMENTS
22
23
SEARCHING OF LINE ART
SEARCHING OF TEXT
SEARCHING OF IMAGES
NORMALISATION OF LINE ART SEARCH RESULTS
NORMALISATION OF TEXT SEARCH RESULTS
NORMALISATION OF IMAGE SEARCH RESULTS
25
COMBINATION OF ALL SEARCH RESULTS
26
RETURN OF COMBINED SEARCH RESULT
27

Figure 2

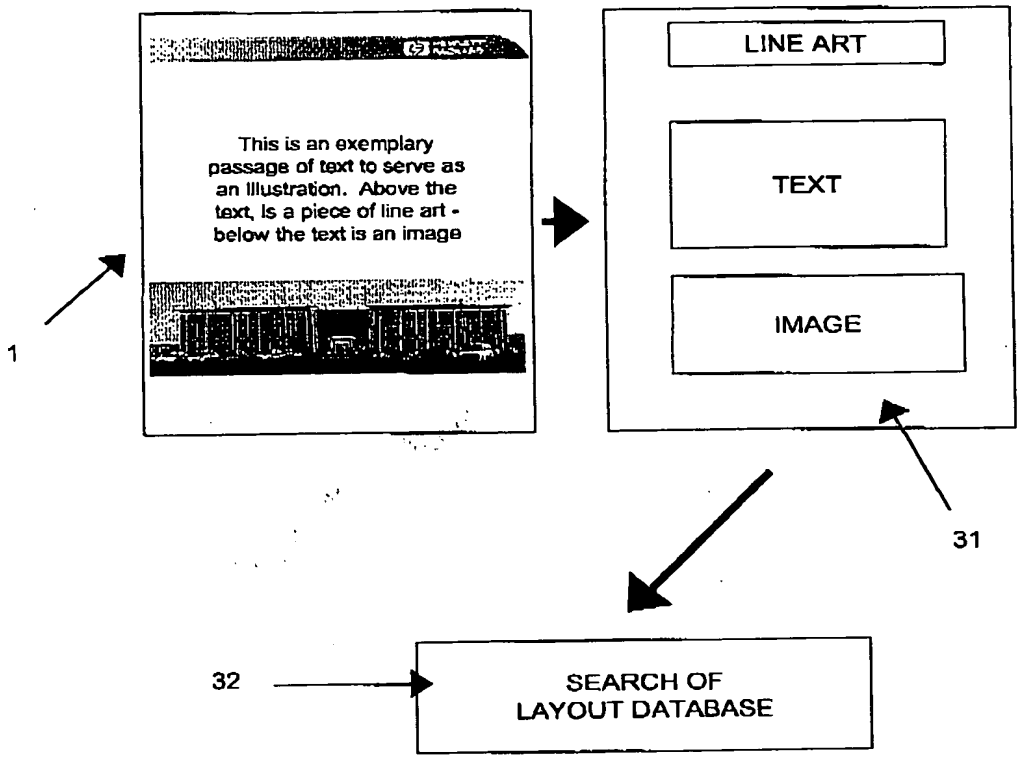
This is an exemplary passage of text to serve as an illustration. Above the text, is a piece of line art - below the text is an image
1
LINE ART
TEXT
IMAGE
31
32
SEARCH OF LAYOUT DATABASE

Figure 3

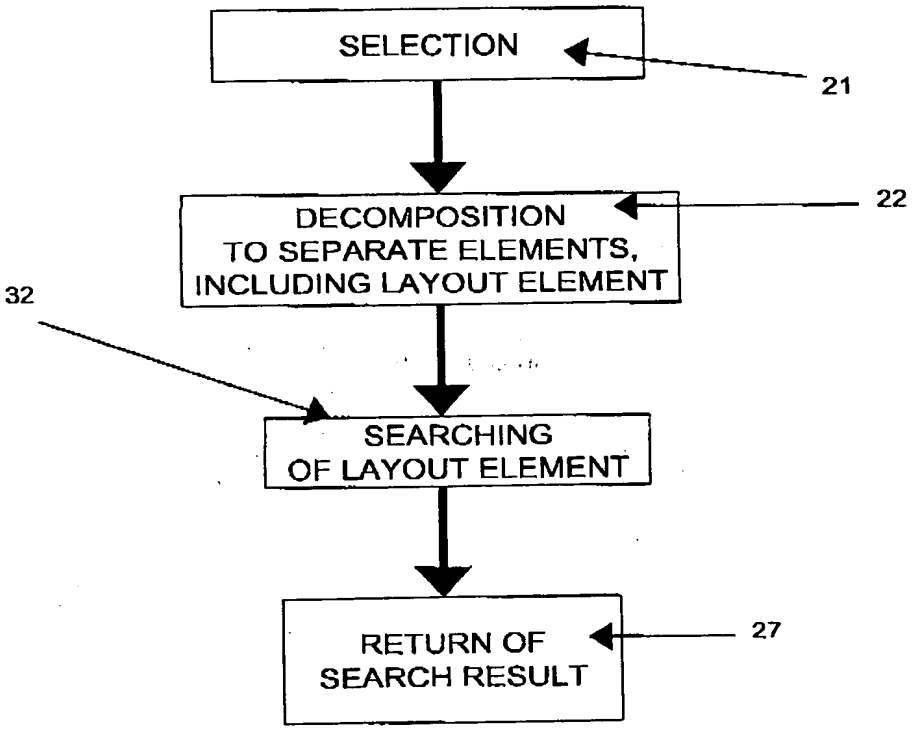
SELECTION
21
DECOMPOSITION
TO SEPARATE ELEMENTS,
INCLUDING LAYOUT ELEMENT
22
32
SEARCHING
OF LAYOUT ELEMENT
RETURN OF
SEARCH RESULT
27

Figure 4

THIS PAGE BLANK (USP